Algo Econ Project - Linear Regression from Strategic Data Sources

Yanjin Chen, Xi Wang, Lanyin Zhang, Tijun Fu

May 2023

1 Introduction

An insurance company is trying to design a linear regression function $y = w^T x + b$, with parameter $w \in \mathcal{R}^d$, $b \in \mathcal{R}$, to determine the insurance payment/premium y for any customer with feature vector $x \in \mathcal{R}^d$.

True feature vector x includes the true values of a customer's age, sex, bmi, children, and etc. However, the insurance company would charge higher premium to riskier customers, which would lead the customer to misreport the feature as some $z \in \mathcal{R}^d$ in order to lower the payment $y = w^T z + b$. The cost of manipulation of misreporting the personal feature would cause a cost $c(z; x)$.

Therefore, we define the optimal feature z^* that a customer with true feature x would misreport to be:

$$z^*(x) = \arg\min_z [w^T z + b + c(z; x)]$$

We say a regressor (w, b) is incentive compatible if $z^*(x) = x$ for any x.

2 Incentive compatible regressors with different cost functions

2.1 Standard Euclidean distance

Suppose $c(z; x) = \sqrt{\sum_{i=1}^d (z_i - x_i)^2}$, which is the standard Euclidean distance to be the cost of misreporting. Since customers are motivated to deceive the insurance company, the costs of fraud should be lower than added profits. And therefore, customers should follow the condition: $w^T(x - z) - c(z; x) > 0$. We propose the following theorem:

Theorem 2.1. *A linear regressor with standard Euclidean distance is incentive compatible if and only if* $||w||_2 \leq 1$.

Proof. We show the forward direction. If $||w||_2 \leq 1$, then we want to show that for all $x, z \in \mathcal{R}^d$, we have

$$w^T x + b \leq w^T z + b + ||z - x||_2$$

This is equivalent to show

$$w^T(x - z) \leq ||z - x||_2$$

Since $z, x \in \mathcal{R}$ are chosen arbitrarily, then we can let $y \in \mathcal{R}^d = x - z$, then the statement becomes

$$w^T y \leq ||y||_2$$

Since $||w||_2 \leq 1$, then by Cauchy-Schwarz inequality we have

$$w^T y \leq ||w||_2 ||y||_2 \leq ||y||_2$$

Hence we finished the forward direction.

Now we show the backward direction that if we have a regressor such that for all $x, z \in \mathcal{R}^d$, we have

$$w^T x + b \leq w^T z + b + ||z - x||_2$$

then $||w||_2 \leq 1$. The problem again can be solved by replacing $y = z - x$, then we need to show

$$w^T y \leq ||y||_2 \Rightarrow ||w||_2 \leq 1$$

Let $y = w$, then $w^T w \leq ||w||_2$, then $||w||_2^2 \leq ||w||_2$, then $||w||_2 \leq 1$. Thus completes the proof. □

2.2 Squared Euclidean distance

Suppose $c(z; x) = \sum_{i=1}^d (z_i - x_i)^2$ is the squared Euclidean distance, which is the cost for pretending to be z. To find the set of all incentive compatible regressors, it must satisfy the following:

Theorem 2.2. *A linear regressor with Squared Euclidean distance is incentive compatible if and only if it is trivial.*

Proof. We only show the forward direction, the backward direction is self-evident. Suppose such incentive compatible regressor w exists, then we have: $x = z^*(x) = \arg\min_z (w^T z + b + \sum_{i=1}^d (z_i - x_i)^2)$
To find w such that $z^*(x) = x$, $\forall x$, fix w, x, and we solve for $z^*(x)$ by applying differentiation:

$$\begin{aligned} \frac{\partial w^T z + b + \sum_{i=1}^d (z_i - x_i)^2}{\partial z} &= (w_1, ..., w_d) + (2(z_1 - x_1), ..., 2(z_d - x_d)) \\ &= (w_1 + 2(z_1 - x_1), ..., w_d + 2(z_d - x_d)) \end{aligned}$$

Since $w^T z + b + c(x, z)$ is convex, then when $z_i = x_i - \frac{w_i}{2} = 0$ it reaches minimum value, then we have:

$$w_i + 2(z_i - x_i) = 0 \Rightarrow \frac{w_i}{2} + z_i - x_i = 0 \Rightarrow z_i = x_i - \frac{w_i}{2}$$

Hence by assumption $z_i = x_i$, then it follows that $x_i = x_i - \frac{w_i}{2}$, yielding $w = 0$. □

3 Convex Optimization Problem Formulation

3.1 Problem Setup

Suppose there are n un-manipulated data $(x_1, y_1), ..., (x_n, y_n)$, where x_i is the true feature and $y_i \in \mathcal{R}$ is the payment of customer i. The strategic customers have a squared Euclidean distance manipulation cost:

$$c(z; x) = \sum_{i=1}^{d} (z_i - x_i)^2$$

The customers' incentive is to lower the cost of their payment and thus has the strategy to choose a fake feature z based on their true feature x:

$$z^*(x) = \arg\min_z [w^T z + b + c(z; x)]$$

On the other hand, for insurance company, empirical risks appears when customers misreports their features z_i instead of x_i. Inspired by the usual loss function of regular linear regression loss

$$l(w) = \sum_{i=1}^{n} (w^T x_i + b - y_i)^2$$

considering that the future customer would manipulate their x_i to fool the regressor, we propose the updated risk for the insurance company to be:

$$l'(w) = \sum_{i=1}^{n} (w^T z_i + b - y_i)^2$$

where $z = z^*(w, x)$.

Another perspective We may also proceed with regular linear regression with loss function $l(w) = \sum_{i=1}^{n} (w^T x_i + b - y_i)^2$, this loss function is universally known as convex. However, when we encounter subsequent customers with fake feature vector z_i, we may use the fact that $z^*(x_i) = x_i - \frac{w}{2}$ to recover their true feature $x_{i_{pred}}$. Then we claim that we can minimize the loss of $l'(w) = \sum_{i=1}^{n} (w^T x_{i_{pred}} + b - y_i)^2$.

3.2 Proof of Convexity

Theorem 3.1. *The loss function $l(w) = \sum_{i=1}^{n} (w^T z_i + b - y_i)^2$ is convex in 1-dimension with the constraint of $b \leq -\frac{x_i^2}{2} + y_i$.*

Proof. For the 1-dimensional case, we have

$$\begin{aligned} l(w) &= \sum_{i=1}^{n} (w^T z_i + b - y_i)^2 \\ &= \sum_{i=1}^{n} (w^T (x_i - \frac{w_i}{2}) + b - y_i)^2 \qquad \text{plug in value of } z_i \text{ from section 2.2} \end{aligned}$$

We take the first derivative w.r.t. the loss function proposed above.

$$\begin{aligned}\frac{dl(w)}{dw} &= \frac{d(w^T(x-\frac{w}{2})+b-y)^2}{dw} \\ &= \frac{d(w^Tx-\frac{w^Tw}{2})+b-y)^2}{dw} \\ &= 2(w^Tx-\frac{w^Tw}{2}+b-y)(x-w)\end{aligned}$$

Then, we take the second derivative w.r.t. the loss function.

$$\begin{aligned}\frac{d^2l(w)}{dw^2} &= \frac{d2(w^Tx-\frac{w^Tw}{2}+b-y)(x-w)}{dw} \\ &= 2(x-w)(x-w)-2(w^Tx-\frac{w^Tw}{2}+b-y) \\ &= 2x^2-4xw+2w^2-2w^Tx+w^Tw-2b+2y \\ &= 2x^2-6wx+3w^2+2y-2b\end{aligned}$$

To guarantee $2x^2-6wx+3w^2+2y-2b \geq 0$, we differentiate it w.r.t. w to get the local minimum and make the minimum value non-negative.

$$\begin{aligned}\frac{d2x^2-6wx+3w^2+2y-2b}{dw} &= -6x+6w=0 \\ w^* &= x\end{aligned}$$

Plug $w^*=x$ into the original equation to get the minimum value: $-x^2+2y-2b$.
$-x^2+2y-2b \geq 0 \Rightarrow b \leq \frac{-x^2+2y}{2}$
Therefore, $b \leq \frac{-x^2+2y}{2}$ can guarantee the second derivative of the loss function to be larger than 0. That is, the loss function $l(w)=\sum_{i=1}^{n}(w^Tz_i+b-y_i)^2$ is convex for $b \leq \frac{-x^2+2y}{2}$.

Attempts for higher dimensions In order to prove convexity in higher dimensions, we need to compute the Hessian matrix and show that it is positive semi-definite. For now we are not equipped with enough mathematical tools to show the work but this can be a future direction. □

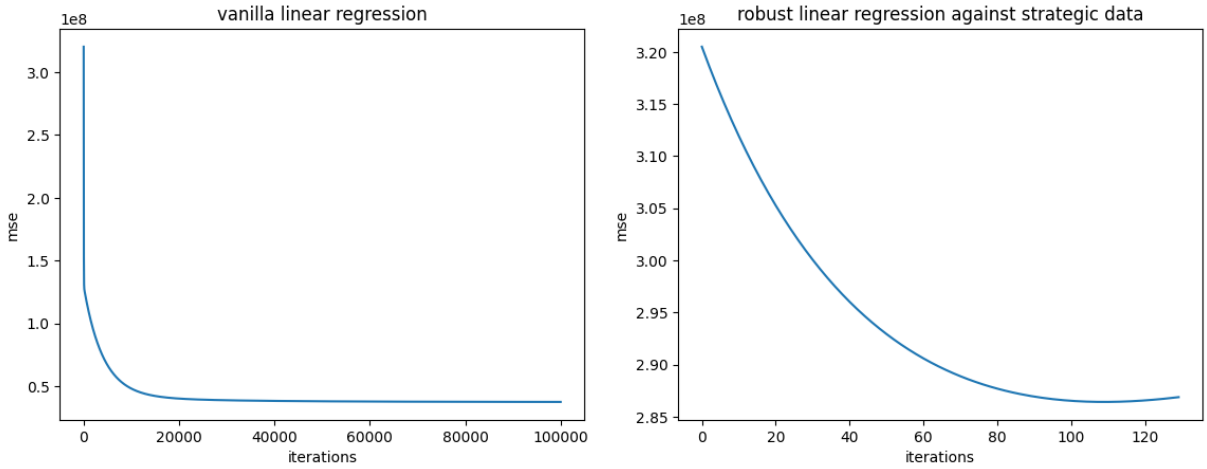
4 Training robust regressor with real world dataset

4.1 Regression task with squared Euclidean cost

We used our loss function derived in the previous section and trained vanilla regression and robust regression on this dataset. The dataset sources can be found at https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction.

We compared our results of vanilla regression versus robust regression against strategic data sources. The process are the same except for different loss function. The gradient descent works vanilla regression as expected, as we see in the mse loss graph, the loss drops and stay plateaued as we proceed in more iterations. However, as shown in the loss graph for strategic data sources, the loss drops in the first 100 iterations and

begin to climb back, suggesting that our loss function may not work well with higher dimensional data.



Regression task with standard Euclidean cost We were not able to find a closed form for z when we are working with standard Euclidean cost, hence the results were not shown for this part.

4.2 Future work: Long-term behavior of the regressors

The regressors trained above is only evaluated in the short-term on the unmanipulated dataset, and as a insurance company we are truly interested in future customers who have the motivation to deviate from his or her true feature x. In fact, if we consider only the squared Euclidean cost function, it is true that for any regressor customers should have incentive to deviate. We formulate what a good robust regressor should have as its properties:

1. w, b should not deviate from the true w_{true}, b_{true} too much, where the distance can be Euclidean or squared Euclidean distance.

2. It should produce minimum regret, where the regret is $(w^T z + b - (w_{true}^T x + b_{true}))^2$

In this way we can evaluate the long-term effect of the regressors, note that empirically w_{true} may be viewed as the regressor produced by vanilla regression.